DOCUMENT RESUME

ED 109 247 TM 004 711

AUTHOR Timm, Neil H.; Carlson, James E.
TITLE Part and Bipartial Canonical Correlation Analysis.
PUB DATE [Apr 75]
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS Computer Programs; *Correlation; Data Analysis; *Hypothesis Testing; *Matrices; *Statistical Analysis; *Tests of Significance
IDENTIFIERS Canonical Correlation Analysis

ABSTRACT

Part and bi-partial canonical correlations were developed by extending the definitions of part and bi-partial correlation to sets of variates. These coefficients may be used to help researchers explore relationships which exist among several sets of normally distributed variates. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED109247

Part and Bipartial Canonical Correlation Analysis

Neil H. Timm and James E. Carlson

University of Pittsburgh

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 004 211

Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April, 1975.

Part and Bipartial Canonical Correlation Analysis

Neil H. Timm and James E. Carlson

University of Pittsburgh

1. Introduction

The concept of simple correlation was introduced into statistics by Sir Francis Galton in several papers published during the 1880's. However, his ideas on correlation were generally unknown until his book, Natural Inheritance, was published in 1889. Galton's work stimulated Pearson (1896, 1898) to develop a precise mathematical theory of correlation which led to the development of partial and multiple correlation (Yule, 1897, 1907). It was not until 1926 that M. Ezekiel and B. B. Smith defined part correlation and, although not explicitly, the notion of bipartial correlation (Ezekiel, 1941).

Although multiple correlation coefficients enable us to investigate associations between one variate and a set of variates, simple, partial, part, and bipartial correlation coefficients are used as measures of association between two variates. Generalizing the notion of correlation between one variate and a set of variates to two sets of variates, Hotelling (1935, 1936) developed canonical correlation coefficients and canonical variates to investigate linear relationships between two sets of variates. However, it was Roy (1957, p. 26) and more recently Rao (1969) who generalized the concept of canonical correlation to partial canonical correlation which is no more than the canonical correlation between two sets of variates $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ after the effect of a third set of variates $\underset{\sim}{Z}$ is removed.

By extending the definitions of part and bipartial correlation to sets of variates, we develop part and bipartial canonical correlations and illustrate how these coefficients and their corresponding canonical variates may be used to explore relationships which exist among sets of normally distributed variates.

2. Canonical Correlation Analysis

Given a set of p ability variates $\underset{\sim}{Y}$ and a set of q personality variates $\underset{\sim}{X}$, where $\underset{\sim}{W}' = [\, \underset{\sim}{Y}' \; \underset{\sim}{X}' \,]$ is normally distributed with variance-covariance matrix Σ,

$$\underset{\sim}{W} = \begin{bmatrix} \underset{\sim}{Y} \\ \underset{\sim}{X} \end{bmatrix} \sim N_{p+q} \left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right\}$$

a researcher may want to assess the degree of relationship between the two sets of variates $\underset{\sim}{Y}$ and $\underset{\sim}{X}$. The method of canonical correlation analysis, developed by Hotelling (1935, 1936) for this purpose, was to determine linear combinations of the original variates, $U = \underset{\sim}{a}'\underset{\sim}{Y}$ and $V = \underset{\sim}{b}'\underset{\sim}{X}$, of unit variance such that the simple correlation between U and V was maximal. The mathematical procedure for accomplishing this is to maximize

$$F_{UV} = \max_{\underset{\sim}{a},\underset{\sim}{b}} \; \underset{\sim}{a}'\Sigma_{12}\underset{\sim}{b}$$

subject to the constraints that $\underset{\sim}{a}'\Sigma_{11}\underset{\sim}{a} = \underset{\sim}{b}'\Sigma_{22}\underset{\sim}{b} = 1$. This leads to the determinantal equation in ρ

$$| \; \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \rho^2 \Sigma_{11} \; | = 0$$

(see for example, Timm, 1975, p. 349). The $s = \min(p,q)$ nonzero positive square roots ρ_i of the roots ρ_i^2 are called the canonical correlations between the canonical variates $U_i = \underset{\sim}{a}_i'\underset{\sim}{Y}$ and $V_i = \underset{\sim}{b}_i'\underset{\sim}{X}$, i=1, 2, ..., s. The coefficient vectors $\underset{\sim}{a}_i$ of the canonical variates U_i are the eigenvectors of the determinantal equation; to obtain the coefficient vectors for each V_i, the relationship

$$\underset{\sim}{b}_i = \frac{\Sigma_{22}^{-1}\Sigma_{22}\,\underset{\sim}{a}_i}{\rho_i} \qquad i=1, 2, \ldots, s$$

is used. The canonical variates within each set, U_i and V_i, are clearly uncorrelated and have unit variance,

$$\text{cov}(U_i, U_j) = \text{cov}(V_i, V_j) = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$$

Furthermore, the covariance between V_i and U_i is ρ_i for i=1, ..., s, and 0, otherwise:

$$\text{cov}(U_i, V_i) = \rho_i \qquad i=1, \ldots, s$$

$$\text{cov}(U_i, V_j) = 0 \qquad i\neq j$$

Investigating the canonical variates further, it is of interest to determine the correlation of each canonical variate for a set with the individual variates within the set. These correlations represent the contribution of each variate to the composite canonical variate and help in the interpretation of canonical variates. The correlations are given by the expressions:

$$\text{corr } (\underset{\sim}{Y}, U_i) = \text{corr } (\underset{\sim}{Y}, \underset{\sim}{a}_i'\underset{\sim}{Y})$$

$$= \Sigma_{11}\, \underset{\sim}{a}_i/\sigma_{y_i}$$

$$\text{corr } (X, V_i) = \Sigma_{22}\, \underset{\sim}{b}_i/\sigma_{x_i}$$

Besides using correlations within a set to better understand canonical variates, we should also examine the relationships between canonical variates in one set and individual variates in the other. These become

$$\text{corr } (\underset{\sim}{X}, U_i) = \text{corr } (\underset{\sim}{X}, \underset{\sim}{a}_i'\underset{\sim}{X})$$

$$= \Sigma_{21}\, \underset{\sim}{a}_i/\sigma_{x_i}$$

$$= \rho_i\, \Sigma_{22}\, \underset{\sim}{b}_i/\sigma_{x_i}$$

$$= \rho_i\, \text{corr } (\underset{\sim}{X}, V_i)$$

$$\text{corr } (Y, V_i) = \rho_i\, \text{corr } (\underset{\sim}{Y}, U_i)$$

To apply the theory developed above for a sample $\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \ldots, \underset{\sim}{Y}_N \sim$ IN $(\underset{\sim}{\mu}, \Sigma)$, the population variance-covariance matrices are replaced by their sample counterparts S_{ij}. Alternatively, sample correlation matrices R_{ij} may also be used since the roots of $S_{11}^{-1/2}\, S_{12}\, S_{22}^{-1}\, S_{21}\, S_{11}^{-1/2}$ and $R_{11}^{-1/2}\, R_{12}\, R_{22}^{-1}\, R_{21}\, R_{11}^{-1/2}$ are identical. Although the vectors $\underset{\sim}{a}_i$ and $\underset{\sim}{b}_i$ associated with the determinantal equation with S_{ij}'s replacing Σ_{ij}'s, will have units of measurement proportional to the original variates, the units of U_i and V_i may not be meaningful. Canonical variates obtained by using correlation matrices have no units of measurement and should be evaluated in terms of standardized variates.

To test the null hypothesis that the p-variates are unrelated to the q-variates

$$H_o: \ \Sigma_{12} = 0$$

$$H_1: \ \Sigma_{12} \neq 0$$

several multivariate criteria have been proposed. Bartlett (1938) outlines a procedure for testing the hypothesis when the sample sizes are large. He defines

$$\Lambda = \prod_{i=1}^{s} (1-r_i^2)$$

where r_i^2 is the sample estimator of ρ_i^2 and uses a chi-square approximation for the distribution of Λ. The hypothesis of independence is rejected if

$$X_B^2 = -[(N-1) - (p + q + 1)/2] \quad \log \Lambda > \chi_\alpha^2 (pq)$$

where χ_α^2 (pq) is the upper α percentile of a chi-squared distribution with pq degrees of freedom.

If the null hypothesis of no relationship (or independence) can be rejected, the contribution of the first root of Λ may be removed and the significance of the remaining roots evaluated (see Bartlett, 1951, or Rao, 1952, p. 370). In general, with $s' < s = \min(p, q)$ roots removed, we define

$$\Lambda^* = \prod_{i=s'+1}^{s} (1-r_i^2)$$

Partitioning Bartlett's chi-square statistic,

$$X_B^2 = -[(N-1) - (p + q + 1)/2] \quad \log \Lambda^*$$

we find that X_B^2 has an approximate chi-square distribution with (p-s')(q-s') degrees of freedom and may be used to test the significance of the roots s' + 1 through s. The tests for significant canonical correlations, other than the first, are very conservative unless the correlations removed are close to 1 (see Williams, 1967).

An alternative to Bartlett's procedure has been developed by Roy (1953) and is called Roy's largest root criterion. To test the significance of each root using Roy's procedure, the parameters

$$s = \min (p-i+1,\ q-i+1)$$

$$m = \frac{|\ p-q\ |-1}{2}$$

$$n = \frac{N - p - q - 2}{2}$$

are defined for Heck's (1960) charts and the characteristic roots r_i^2 are compared to a critical value $\theta^{\alpha}(s,m,n)$, found in the appropriate chart.

The hypothesis of independence for two sets of random variates reduces to some familiar univariate results. If the number of variates in the p set is one, the hypothesis reduces to

$$H_o: \rho_{12} = \rho \qquad \rho^2_{o(1,\ \ldots,\ q)} = 0$$

or

$$H_1: \rho_{12} \neq \rho \qquad \rho^2_{o(1,\ 2,\ \ldots,\ q)} \neq 0$$

and is tested using $F = (R^2/q)/[(1-R^2)/(N-q-1)]$ where R^2 is the square of the sample multiple correlation coefficient. For $p = q = 1$, the hypothesis reduces to

$$H_o: \quad \rho = 0$$

$$H_1: \quad \rho \neq 0$$

and is tested using $t = r\sqrt{N-2}\ /\ \sqrt{1-r^2}$ where r is the sample correlation coefficient (Fisher, 1915).

3. Partial Canonical Correlation Analysis

Extending Hotelling's development of canonical analysis to three sets of variates, Rao (1969) using some results given in Anderson (1958, p. 33) developed the notion of partial canonical correlation analysis which may be used to assess the partial independence of two sets of variates given a third set of variates.

Given a set of p variates $\underset{\sim}{Y}$, a set of q variates $\underset{\sim}{X}$ and a set of r variates $\underset{\sim}{Z}$, where $\underset{\sim}{U}' = [\underset{\sim}{Y}, \underset{\sim}{X}, \underset{\sim}{Z}]$ is normally distributed,

$$\underset{\sim}{U} = \begin{bmatrix} \underset{\sim}{Y} \\ \underset{\sim}{X} \\ \underset{\sim}{Z} \end{bmatrix} \sim N_{p+q+r} \left\{ \begin{bmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \\ \underset{\sim}{\mu}_3 \end{bmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix} \right\}$$

we may be interested in assessing the degree of the relationship between $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ after removing the linear effect of the variates in the $\underset{\sim}{Z}$ set. That is, we want to find from the variates $\underset{\sim}{e}_Y = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}}$ and $\underset{\sim}{e}_X = \underset{\sim}{X} - \hat{\underset{\sim}{X}}$ where $\underset{\sim}{e}_X$ and $\underset{\sim}{e}_Y$ are the residual vectors (obtained from regressing $\underset{\sim}{X}$ on $\underset{\sim}{Z}$ and $\underset{\sim}{Y}$ on $\underset{\sim}{Z}$) linear combinations, $U = \underset{\sim}{a}'\underset{\sim}{e}_Y$ and $V = \underset{\sim}{b}'\underset{\sim}{e}_X$, of unit variance such that the simple correlation between U and V is maximal. Mathematically, this is equivalent to maximizing

$$F_{UV} = \max_{\underset{\sim}{a},\underset{\sim}{b}} \underset{\sim}{a}' \Sigma_{12.3} \underset{\sim}{b}$$

subject to the constraints $\underset{\sim}{a}' \Sigma_{11.3} \underset{\sim}{a} = \underset{\sim}{b}' \Sigma_{22.3} \underset{\sim}{b} = 1$. The matrices $\Sigma_{ij.3}$ are the elements of the variance-covariance matrix of the residual vectors $\underset{\sim}{e}_Y$ and $\underset{\sim}{e}_X$.

$$\Sigma_{.3} = \begin{pmatrix} \Sigma_{11.3} & \Sigma_{12.3} \\ \Sigma_{21.3} & \Sigma_{22.3} \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11} - \Sigma_{13}\,\Sigma_{33}^{-1}\,\Sigma_{31} & \Sigma_{12} - \Sigma_{13}\,\Sigma_{33}^{-1}\,\Sigma_{32} \\ \Sigma_{21} - \Sigma_{23}\,\Sigma_{33}^{-1}\,\Sigma_{31} & \Sigma_{22} - \Sigma_{23}\,\Sigma_{33}^{-1}\,\Sigma_{32} \end{pmatrix}$$

the variance-covariance matrix of the conditional distribution of $\underset{\sim}{Y}$ and $\underset{\sim}{X}$, given $\underset{\sim}{Z}$.

Maximizing F_{UV}, as shown by Rao (1969), leads to the determinantal equation in ρ

$$| \Sigma_{12.3}\,\Sigma_{22.3}^{-1}\,\Sigma_{21.3} - \rho_{.3}^{2}\,\Sigma_{11.3} | = 0$$

The $s = \min(p, q)$ nonzero positive square roots $\rho_{i.3}$ of the roots $\rho_{i.3}^{2}$ are called the partial canonical correlations between the partial canonical variates $U_i = \underset{\sim}{a}_i' \underset{\sim}{Y}$ and $V_i = \underset{\sim}{b}_i' \underset{\sim}{X}$, $i=1, \ldots, s$. The coefficient vectors of each U_i are the eigenvectors of the determinantal equation and the relationship between $\underset{\sim}{a}_i$ and $\underset{\sim}{b}_i$ is given by

$$\underset{\sim}{b}_i = \frac{\Sigma_{22.3}^{-1}\,\Sigma_{21.3}\,\underset{\sim}{a}_i}{\rho_{i.3}} \qquad i = 1, \ldots, s$$

To test the hypothesis of partial independence,

$$H_o: \Sigma_{12.3} = 0$$

$$H_1: \Sigma_{12.3} \neq 0$$

using Roy's criterion, the parameters are defined by

$$s = \min(p - i + 1, q - i + 1)$$

$$m = \frac{|p - q| - 1}{2}$$

$$n = \frac{N - r - p - q - 2}{2}$$

and the $r^2_{i.3}$ are compared to the critical value θ^{α} (s, m, n) at the level α found from the Heck charts. Alternatively, defining Λ as

$$\Lambda = \prod_{i=1}^{s} (1-r^2_{i.3})$$

Bartlett's criterion

$$X^2_B = -[(N-r-1) - (p + q + 1)/2] \log \Lambda \sim \chi^2 (pq)$$

may be used.

Some familiar univariate results are evident from the multivariate procedure. If p = 1, the partial canonical correlation coefficient becomes the partial multiple correlation coefficient (see Rao, 1973, p. 268). Setting p = q = 1, the test of partial independence reduces to testing

$$H_o: \rho_{12.3} = 0$$

$$H_1: \rho_{12.3} \neq 0$$

which is tested using the familiar t statistic, $t = r_{12.3}\sqrt{N-3} \,/\, \sqrt{1-r^2_{12.3}}$ where $r_{12.3}$ is the sample partial correlation coefficient (see Anderson, 1958, p. 84),

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r^2_{13}}\ \sqrt{1-r^2_{23}}}$$

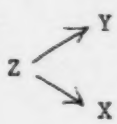
Following Fisher (1924), to test that a partial correlation coefficient is equal to zero under normality, the t statistic for testing that a simple correlation coefficient is equal to zero is modified by subtracting one degree of freedom from the degrees of freedom for error for every variate removed and the simple correlation coefficient is replaced by a partial correlation coefficient. Extending this rule to the partial canonical

correlation analysis procedure, we were able to obtain, by analogue, a test for multivariate partial independence.

4. Part and Bipartial Canonical Correlation Analysis

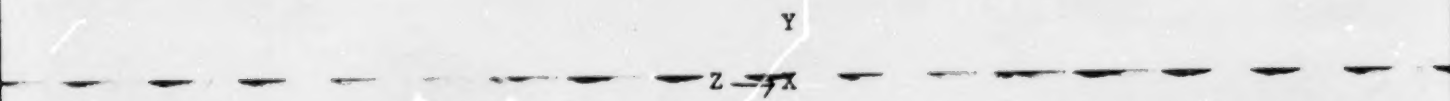
Following Pearson, partial correlation coefficients under normality are no more than simple correlation coefficients in conditional distributions. Holding several variates constant in a multivariate normal distribution allows one to investigate relationships between two variates while controlling for the other variates which directly influence the two variates under study. Such an explanation of a partial correlation coefficient would not satisfy most researchers. Alternatively we find people saying that a simple partial correlation coefficient represents an estimate of what a simple correlation would be if we were able to calculate the simple correlation coefficient at any one of several levels of a third fixed variable. This is still unsatisfactory since we never check this when we use the coefficient.

Going back to the derivation of a partial correlation coefficient, we said that it is the correlation in residuals after linear effects due to a common variate or set of variates is removed. Implied in this statement is the following causal relationship given by the causal system:



If Z does not influence the variation in X and Y as shown above, the interpretation of a partial correlation coefficient is unclear since by "partralling out" Z from X and Y we are removing the linear effect of Z on both X and Y

Provided Z influences both X and Y, interpretation of the partial correlation coefficient is meaningful. It would not make sense to calculate a partial correlation coefficient if Z influences X but not Y. For this situation we would have the following diagram:



For this case the correlation between X and Y is best estimated by controlling for the influence of Z on X. That is, we want the correlation between Y and X partialling out Z from X and not Y. Such a correlation coefficient is called a part correlation coefficient and for the three variate case is represented by

$$\rho_{1(2.3)} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1 - \rho_{23}^2}}$$

To derive $\rho_{1(2.3)}$ following Yule, we assume a linear relationship between X and Z, $X = \alpha + \beta Z$, and find the simple correlation between $\varepsilon = X - \alpha - \beta Z$ and Y.

As shown by McNemar (1969, p. 322) the test statistic for testing

$$H_o: \rho_{1\,(2.3)} = 0$$

$$H_1: \rho_{1(2.3)} \neq 0$$

under joint normality is $t = r_{1(2.3)}\sqrt{N - 3} \,/\, \sqrt{1-r_{12.3}^2}$. Unfortunately, one may not merely substitute part correlations for partial correlations in the formula for testing $\rho_{12.3} = 0$ to test that a part correlation coefficient is equal to zero. Since $\rho_{1(2.3)}^2 \leq \rho_{12.3}^2$, as may be seen by examining the formulas for these two coefficients, substituting $r_{1(2.3)}^2$ for $r_{12.3}^2$ in the t statistic for testing $\rho_{1(2.3)} = 0$ yields only an approximate test procedure for testing part independence.

To extend the notion of part correlation to the multivariate case, we again assume that

$$\underset{\sim}{U} = \begin{bmatrix} \underset{\sim}{Y} \\ \underset{\sim}{X} \\ \underset{\sim}{Z} \end{bmatrix} \sim N_{p+q+r} \left\{ \begin{pmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \\ \underset{\sim}{\mu}_3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix} \right\}$$

Now however, we are interested in assessing the degree of relationship between $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ after removing the linear effect of the variates in the $\underset{\sim}{Z}$ set from $\underset{\sim}{X}$ and not $\underset{\sim}{Y}$. That is, we want to find linear combinations of the variates $\underset{\sim}{e}_X$ and $\underset{\sim}{Y}$, $U = \underset{\sim}{a}'\underset{\sim}{Y}$ and $V = \underset{\sim}{b}'\underset{\sim}{e}_X$, of unit variance such that the correlation between U and V is maximal. This is equivalent to maximizing

$$F_{UV} = \max_{a,b} \; \underset{\sim}{a}' \; \Sigma_{1(2.3)} \; \underset{\sim}{b}$$

subject to the constraints $\underset{\sim}{a}' \; \Sigma_{11} \; \underset{\sim}{a} = 1$ and $\underset{\sim}{b}' \; \Sigma_{22.3} \; \underset{\sim}{b} = 1$ where the matrix $\Sigma_{1(2.3)}$ is defined by

$$\Sigma_{1(2.3)} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12.3} \\ \Sigma_{21.3} & \Sigma_{22.3} \end{pmatrix}$$

This again leads us to finding the roots and vectors of a determinantal equation,

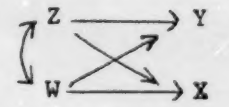
$$| \; \Sigma_{12.3} \; \Sigma_{22.3}^{-1} \; \Sigma_{21.3} - \rho^2_{1(2.3)} \; \Sigma_{11} \; | = 0$$

In addition to part and partial correlation coefficients, other simple correlations are also important to the understanding of linear relationships among variates; for example, suppose two variates Z and W are highly correlated and that the causal relationship among four variates is as follows:

To determine the correlation between Y and X in this case, the linear influence of Z on Y and of W on X is controlled by removing the influence of Z on Y and of W on X. This leads us to the bipartial correlation coefficient

$$\rho_{(1.3)(2.4)} = \frac{\rho_{12} - \rho_{13}\rho_{23} - \rho_{14}\rho_{24} + \rho_{13}\rho_{34}\rho_{24}}{\sqrt{1 - \rho_{13}^2}\quad\sqrt{1 - \rho_{24}^2}}$$

which reduces to a part correlation coefficient if either ρ_{13} or ρ_{24} equals zero. Alternatively if the relationship among the variates were given by the system



the partial correlation coefficient

$$\rho_{12.34} = \frac{\rho_{12.3} - \rho_{14.3}\rho_{24.3}}{\sqrt{1 - \rho_{14.3}^2}\ \sqrt{1 - \rho_{24.3}^2}}$$

would be of interest. The causal relationships among variates influences the researcher's selection of a correlation coefficient and hence the analysis of a set of data.

To extend the idea of a bipartial correlation coefficient to four sets of variates we assume that

$$\underset{\sim}{U} = \begin{bmatrix} \underset{\sim}{Y} \\ \underset{\sim}{X} \\ \underset{\sim}{Z} \\ \underset{\sim}{W} \end{bmatrix} \sim N_{p+q+r+t}\left\{ \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{pmatrix} \right\}$$

and letting

$$\Sigma_{12}^{*} = \Sigma_{12} - \Sigma_{13}\,\Sigma_{33}^{-1}\,\Sigma_{32} - \Sigma_{14}\,\Sigma_{44}^{-1}\,\Sigma_{42} + \Sigma_{13}\,\Sigma_{33}^{-1}\,\Sigma_{34}\,\Sigma_{44}^{-1}\,\Sigma_{42}$$

we form the matrix

$$\Sigma_{(1.3)(2.4)} = \begin{pmatrix} \Sigma_{11.3} & \Sigma_{12}^{*} \\ \Sigma_{21}^{*} & \Sigma_{22.4} \end{pmatrix}$$

which is the variance-covariance matrix of the residuals $\underset{\sim}{e}_Y = Y - \hat{Y}$ and $\underset{\sim}{e}_X = X - \hat{X}$ where $\hat{Y}$ is found by regressing $\underset{\sim}{Y}$ on $\underset{\sim}{Z}$ and $\hat{X}$ is found by regressing $\underset{\sim}{X}$ on $\underset{\sim}{W}$. Notice that if the third set of variates are not in our model that $\Sigma_{(1.3)(2.4)}$ reduces to $\Sigma_{1(2.4)}$ which is analogous to the univariate case.

To assess the degree of relationship between $\underset{\sim}{e}_Y$ and $\underset{\sim}{e}_X$ we again want to maximize the correlation between $U = \underset{\sim}{a}'\underset{\sim}{e}_Y$ and $V = \underset{\sim}{b}'\underset{\sim}{e}_X$. This leads to the solution of the determinantal equation

$$\left| \Sigma_{12}^{*}\ \Sigma_{22.4}^{-1}\ \Sigma_{21}^{*} - \rho^{2}_{(1.3)(2.4)}\ \ \Sigma_{11.3} \right| = 0$$

The $s = \min(p, q)$ nonzero positive square roots $\rho_{i(1.3)(2.4)}$ are the bipartial canonical correlations between the bipartial canonical variates $U_i = \underset{\sim}{a}_i'\,\underset{\sim}{e}_Y$ and $V_i = \underset{\sim}{b}_i'\,\underset{\sim}{e}_X$, $i=1, \ldots, s$. The relationship between the coefficients is given by

$$\underset{\sim}{b}_i = \frac{\Sigma_{22.4}^{-1}\ \Sigma_{21}^{*}\ \underset{\sim}{a}_i}{\rho_{i(1.3)(2.4)}} \qquad i = 1, \ldots, s.$$

To test the hypothesis of bipartial independence

$$H_o: \ \Sigma_{12}^{*} = 0$$

$$H_1: \ \Sigma_{12}^{*} \neq 0$$

we have for this test only an approximate procedure in the multivariate case. The parameters using Roy's criterion are given by

$$s = \min (p - i + 1, q - i + 1)$$

$$m = \frac{|p - q| - 1}{2}$$

$$n = \frac{N - \max (r, t) - p - q - 2}{2}$$

and the bipartial canonical correlations are compared to the critical values found in the appropriate Heck chart.

Defining Λ as

$$\Lambda = \prod_{i=1}^{s} (1 - r^2_{i(1.3)(2.4)})$$

Bartlett's criterion defined by

$$\chi^2_B = -[(N - \max (r, t) - 1) - (p + q + 1)/2] \quad \log \Lambda \sim \chi^2 (pq)$$

might also be used.

The approximate test of part independence for the case of multivariate part canonical analysis follows a similar procedure with r replacing max (r, t) in the formulas and $\Sigma_{12.4}$ replacing Σ^*_{12} in the hypothesis.

5. Example 1: Canonical Correlation Analysis

Suppose a researcher was interested in investigating the relationship between three achievement variables A_1, A_2, and A_3 and two personality variables P_1 and P_2 where the correlation matrix among the variates $Y = \{P_1, P_2\}$ and $X = \{A_1, A_2, A_3\}$ is

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} = \left(\begin{array}{cc|cc|c} 1.0000 & & & & \\ .7951 & 1.0000 & & & \\ \hline .2617 & .3341 & 1.0000 & & \\ .6720 & .5876 & .3703 & 1.0000 & \\ .3390 & .3404 & .2114 & .3548 & 1.000 \end{array}\right.$$

Solving the determinantal equation

$$| R_{12} R_{22}^{-1} R_{21} - \rho^2 R_{11} | = 0$$

the sample canonical correlations r_1 and r_2 are obtained from the roots r_1^2 and r_2^2,

$$r_1^2 = .4746 \qquad r_2^2 = .0375$$

$$r_1 = .6889 \qquad r_2 = .1936$$

Rejecting the hypothesis of independence and finding that only the first root is significantly different from zero, the researcher at first blush might conclude that the two sets of variates are significantly related and that the proportion of variance in common to the two standardized canonical variates

$$U_1 = .7752 \; P_1 + .2662 \; P_2$$

$$V_1 = .0520 \; A_1 + .8991 \; A_2 + .1831 \; A_3$$

is $r_1^2 = .4746$. However, further investigation into the significant canonical variates and the variates within a set would yield that the

$$\text{corr}\,(Z_y, U_1) = R_{11}\,\underset{\sim}{a}_1 = \begin{pmatrix} .987 \\ .883 \end{pmatrix}$$

$$\text{corr}\,(Z_x, V_1) = R_{22}\,\underset{\sim}{b}_1 = \begin{pmatrix} .424 \\ .983 \\ .513 \end{pmatrix}$$

which indicates that P_1 and P_2 are equally important to U_1, but that A_2 is more important to V_1 than either A_1 or A_2. Furthermore,

$$U_Y^2 = \frac{(.987)^2 + (.883)^2}{2} = .876$$

of the variance of the first set is accounted for by U_1 and only

$$V_X^2 = \frac{(.424)^2 + (.983)^2 + (.513)^2}{2} = .470$$

of the variance in the other set is accounted for by V_1. Investigation of the correlations between the variates in one set and the significant canonical variates yields

$$\text{corr } (X, U_1) = \rho_1 \quad \text{corr } (Z_x, V_1) = \begin{pmatrix} .292 \\ .677 \\ .353 \end{pmatrix}$$

$$\text{corr } (Y, V_1) = \rho_1 \quad \text{corr } (Z_y, U_1) = \begin{pmatrix} .680 \\ .608 \end{pmatrix}$$

This shows that A_2 in the achievement set is influenced most by the personality canonical variate and that the achievement canonical variate is influenced equally by both personality variates. More specifically, 22% of the variance common to the achievement variates can be accounted for by a linear combination of personality variables,

$$U_{X|U_1}^2 = \frac{(.292)^2 + (.677)^2 + (.355)^2}{3} = .223$$

whereas the proportion of variance in the personality variables accounted for by the achievement canonical variate is 42%,

$$V^2_{Y|V_1} = \frac{(.680)^2 + (.608)^2}{2} = .416$$

In summary, given the two sets of variates

$$Y = \{P_1, P_2\} \quad \text{and} \quad X = \{A_1, A_2, A_3\}$$

it appears that the proportion of variance "in common" to the two significant canonical variates is about 47%. However, 88% of the variance in the set $\underset{\sim}{Y}$ is accounted for by U_1 and only 42% of the variance in $\underset{\sim}{Y}$ is accounted for by V_1. Similarly, 47% of the variance in the set $\underset{\sim}{X}$ is accounted for by V_1, but only 23% of the variance in $\underset{\sim}{X}$ is accounted for by the canonical variate U_1.

Stewart and Love (1968) observed that

$$V^2_{Y|V_1} = U^2_Y \, r_1^2$$

$$U^2_{X|U_1} = V^2_X \, r_1^2$$

and termed these redundancy indexes since they better summarize the overlap between two sets of variates than the square of a canonical correlation. For our data $V^2_{Y|V_1} = .416$ and $U^2_{X|U_1} = .223$. That is, the redundancy in $\underset{\sim}{Y}$ given $\underset{\sim}{X}$ is .416 and the redundancy in $\underset{\sim}{X}$ given $\underset{\sim}{Y}$ is .226. The larger the redundancy indexes, the larger the overlap among the variates in each domain.

Example 2: Bipartial Canonical Correlation Analysis

Using a random sample of 502 twelfth-grade students from the project. Talent survey (supplied by William W. Cooley at the University of Pittsburgh), data were collected on 11 tests: (1) general information test, part I, (2) general information test, part II, (3) English, (4) reading comprehension, (5) creativity, (6) mechanical reasoning, (7) abstract reasoning, (8) mathematics, (9) sociability inventory, (10) physical science interest inventory, and (11) office work interest inventory. Knowing that verbal ability tests (3) through (5) are highly correlated with the nonverbal tests (6) through (8), the investigator was interested in investigating the relationship between the general information tests (1) and (2) and the interest inventory measures, tests (9) through (11). Since prior knowledge would indicate that the relationships among the sets are of the form

highly correlated

$$Z = \{3, 4, 5\} \longrightarrow Y = \{1, 2\}$$
$$W = \{6, 7, 8\} \longrightarrow X = \{9, 10, 11\}$$

the set of data lends itself to a bipartial canonical analysis. The correlation matrix for the sets of variates is shown in Table 1.

Table 1. Original Variate Intercorrelation Matrix

	Y		X			Z			W		
Y	1.000										
	.861	1.000									
	-.011	.062	1.000								
X	.573	.397	.055	1.000							
	-.349	-.234	.084	-.246	1.000						
	.492	.550	.083	.094	.109	1.000					
Z	.698	.765	.021	.275	-.087	.613	1.000				
	.644	.621	.001	.340	-.119	.418	.595	1.000			
	.661	.519	-.075	.531	-.364	.160	.413	.522	1.000		
W	.487	.469	.007	.202	-.079	.456	.530	.433	.451	1.000	
	.761	.649	.030	.500	-.191	.566	.641	.556	.547	.517	1.000

Using the CANON computer program described in Section 8 we find that the matrix of partial variances and covariances is

$$\left[\begin{array}{c|c} S_{Y\cdot Z} & S_{(Y\cdot Z)(X\cdot W)} \\ \hline S'_{(Y\cdot Z)(X\cdot W)} & S_{X\cdot W} \end{array}\right] = \left[\begin{array}{cc|ccc} .429 & .263 & -.012 & .133 & -.163 \\ .263 & .365 & .051 & .060 & -.110 \\ \hline -.012 & .051 & .987 & .076 & .054 \\ .133 & .060 & .076 & .635 & -.041 \\ -.163 & -.110 & .054 & -.041 & .858 \end{array}\right]$$

and the eigenvalues of the determinantal equation

$$\left| S_{(Y\cdot Z)(X\cdot W)} S^{-1}_{X\cdot W} S'_{(Y\cdot Z)(X\cdot W)} - \rho^2_{(Y\cdot Z)(X\cdot W)} S_{Y\cdot Z} \right| = 0$$

are .133 and .022. Using Roy's criterion for the first root we have s = 2, m = 0, and n = 247.5 and using the Heck (1960) charts we find that this root differs from zero at the .01 level. Similarly for the second root s = 1, m = 0 and n = 247.5. When s = 1 we calculate the F-statistic

$$F = \left(\frac{n+1}{m+1}\right)\left(\frac{r^2}{1-r^2}\right)$$

(see, for example, Morrison, 1967, p. 166-167) and the test statistic distribution is $F_{2m+2,\ 2n+2}$. For our data

$$F = \left(\frac{248.5}{1}\right)\left(\frac{.022}{.978}\right) = 5.590$$

Referring to tables of F we find that the second root also differs from zero at the .01 level.

The bipartial canonical correlation coefficients are .364 and .150. Using Bartlett's approximate chi squared test we find that the hypothesis of bipartial independence is rejected for both roots (chi squared = 81.676, df = 6, $p < .0001$) and also for the second root after having removed the first root (chi squared = 11.223, df = 2, p = .00367). Thus we reach the same conclusions using Bartlett's test as we do using Roy's.

The standardized canonical variates are

$$U_1 = 1.674Y_1 - 0.255Y_2$$

$$U_2 = -1.180Y_1 + 2.205Y_2$$

$$V_1 = -0.120X_1 + 0.863X_2 - 0.737X_3$$

$$V_2 = 0.919X_1 - 0.397X_2 - 0.461X_3$$

and the correlation coefficients between the original and canonical variates are shown in Table 2.

Table 2. Original Variate-Partial Canonical Variate Correlations

	U_1	U_2	V_1	V_2
Y_1	.993	.576	.362	.017
Y_2	.115	.817	.210	.122
X_1	-.034	.128	-.093	.858
X_2	.260	-.031	.715	-.205
X_3	-.265	-.053	-.728	-.356

Examination of these correlation coefficients helps to understand the relationships existing among the original variates and the partial canonical variates. The printout from the CANON program also indicates that U_1 accounts for .66 of the Y-set variance and U_2 accounts for .34. Similarly V_1 accounts for .35 of the X-set variance and V_2 accounts for .30.

The redundancies, or proportions of variance in the Y-set and X-set that are accounted for by the significant canonical variates derived from the opposite sets are shown in Table 3.

Table 3. Redundancies

	V_1	V_2	Overall
Y-set	.087	.008	.095
	U_1	U_2	Overall
X-set	.046	.007	.053

These data indicate that although there are significant relationships between the information tests and interest inventories after partialing out verbal ability and nonverbal ability measures, respectively, the proportions of accounted for variance are rather small. Examining the correlations in Tables 2 and 3 we see that the strongest relationship is between V_1 and the Y-set, and that X_2 and X_3 contribute most to V_1, X_1 being almost uncorrelated with V_1. The next strongest relationship is between U_1 and the X-set, with Y_1 contributing much more to U_1 than does Y_2.

7. The CANON Computer Program

The CANON computer program allows the researcher to analyze multivariate data by any of the four techniques discussed in this paper: Canonical Analysis, Partial Canonical Analysis, Part Canonical Analysis, and Bipartial Canonical Analysis.

The user may input raw data, a variance-covariance matrix, or a correlation matrix, and specifies the type of analysis and number of variates in each set. The first two sets of variates are referred to as the Y-set and the X-set and are the sets whose relationship is to be studied. The third set (Z), if used, contains the variates to be partialed out of the Y-set and the X-set in partial canonical analysis, the Y-set or the X-set in part canonical analysis or the Y-set in bipartial canonical analysis. The fourth set (W) contains the variates to be partialed out of the X-set in bipartial canonical analysis.

The number of variates in the Y-set must be less than or equal to the number of variates in the X-set. Also, the variates must be input in the following order: Y-set, X-set, Z-set, W-set.

The program is written in FORTRAN IV for a DEC PDP10. All calculations are done using double precision. Conversion of the program for other computer systems should not be difficult. Since the program stores "PROBL" and "FINIS" in two single-precision memory locations and checks the first five characters of the title and finish cards with the contents of these locations, changes will be necessary for computers that do not store 5 alphanumeric characters in a single precision memory location. Similar changes will be necessary for some of the labels for the output which are also stored in memory via DATA statements. These changes may be the only changes required for many computers but the user should check that the names of FORTRAN-supplied functions used in the program correspond to those available on the available system. Listings of the programs and card decks are available upon request from the authors.

INPUT TO CANON

The input to the program is as follows:

(a) Title Card

The title card contains the characters PROBL in columns 1 through 5 and any title that the user chooses in columns 6 through 80.

(b) Problem Card

The second card contains 9 numbers specifying the nature of the problem and type of analysis. The first 8 numbers are integers and each is punched in a 5-digit field, right justified. The 9th number is a significance level to be used as a criterion for defining significant canonical relationships, according to Bartlett's test, and is a 4-digit decimal fraction punched with a decimal point. The numbers in this card are:

Col. 1-5 — N = No. of observations in the sample

Col. 6-10 — NP = no. of variates in the Y-set

Col. 11-15 — NQ = No. of variates in the X-set (NP≤NQ)

Col. 16-20 — NR = No. of variates in the Z-set (punch zero or leave blank if no Z-set)

Col. 21-25 — NT = No. of variates in the W-set (punch zero or leave blank if no W-set)

Col. 26-30 — Punch 1 if Canonical Analysis
Punch 2 if Partial Canonical Analysis
Punch 3 if Part Canonical Analysis, Partialing Z-set from Y-set
Punch 4 if Part Canonical Analysis, Partialing Z-set from X-set
Punch 5 if Bipartial Canonical Analysis

Col. 31-35 — NRMC = No. of format cards

Col. 36-40 — Punch 0 or leave blank if raw data to be input
Punch 1 if variance-covariance or correlation matrix to be input

Col. 41-45 — PIN = significance level for retention of canonical variates according to the Bartlett test, Punched with decimal point. Punch 1.0 if it is desired to have all possible canonical variates extracted.

(c) Format Card

The input format contains one F-field for each variate that is input. The user should remember the order in which the variate sets must be input, as specified below.

(d) Data

The data may be input in raw form (IN=zero) or in the form of a variance-covariance or correlation matrix (IN=one).

(i) Raw Data: The values on the variates for each observation are input in a single record containing one or more cards. The order of input must be: Y-set variates, X-set variates, Z-set (if used) variates, W-set (if used) variates. The variates are punched as specified on the variable format card, card c.

(ii) Variance-covariance or Correlation Matrix: The complete square symmetric matrix of variances and covariances or inter-correlations of all variates is input. The matrix must be in the form:

$$\begin{bmatrix} S(Y,Y) & S(Y,X) & S(Y,Z) & S(Y,W) \\ S(X,Y) & S(X,X) & S(X,Z) & S(X,W) \\ S(Z,Y) & S(Z,X) & S(Z,Z) & S(Z,W) \\ S(W,Y) & S(W,X) & S(W,Z) & S(W,W) \end{bmatrix}$$

where S(A,B) represents the variance-covariance matrix or correlation matrix of variate-set A with variate-set B. The number of variates in the Y-set must be less than or equal to the number in the X-set.

The values in each row of the matrix are input in one record containing one or more cards, punched as specified in the variable format card, card c.

(e) End of Job Card

The program allows the user to stack jobs to be run sequentially, each job containing a complete set of cards a through d. Thus if a second job is to be run, a second title, problem, etc. card follows the data from the first job. The data for the last job is followed by a end-of-job card which contains the characters "FINIS" in columns 1 through 5.

OUTPUT FROM CANON

The output from CANON includes the following (all values are printed in scientific notation; eg. .1234 D-01 = .1234 x 10^{-1} = .01239):

(a) Variance-covariance matrix (or correlation matrix when it is input) of all variates.

(b) Standard deviations of all variates, by set

(c) Variance-covariance matrix after partialing. Output when the analysis is a partial, part or bipartial canonical analysis, this matrix contains the variances and covariances of the Y and X sets after partialing.

(d) Eigenvalues from the determinantal equation formed for the analysis and the values necessary for determining significance by Roy's criterion using the Heck charts.

(e) Canonical, Partial canonical, Part canonical or Bipartial canonical correlation coefficients and Bartlett's test for the significance of the coefficients.

(f) Standardized* canonical coefficients for the Y-set variates and correlation coefficients between the Y-set variates and canonical variates derived from the Y-set.

(g) Standardized canonical coefficients for the X-set variates and correlation coefficients between the X-set variates and canonical variates derived from the X-set.

(h) Proportions of variance in the Y-set accounted for by each significant (Bartlett's test) canonical variate derived from the Y-set, and the similar proportions for the X-set.

(i) Correlation coefficients between Y-set variates and the significant canonical variates derived from the X-set, along with the redundancy for each canonical variate and the overall redundancy.

(j) Correlation coefficients between X-set variates and the significant canonical variates derived from the Y-set, along with the redundancies for each canonical variate and the overall redundancy.

---

*Canonical variates normalized to have unit variance in the sample.

8. References

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. New York: John Wiley.

Bartlett (1938). Further aspects of the theory of multiple regression. Proceedings of the Cambridge Philosophical Society, 34, 33-40.

Bartlett (1951). The goodness of fit of a single hypothetical discriminant function in the case of several groups. Annals of Eugenics, 16, 199-214.

Ezekiel, M. (1941). Methods of Correlation Analysis, Second Edition, New York: John Wiley.

Fisher, R. A. (1915). The frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, Biometrika, 10, 507-521.

Fisher, R. A. (1924) The distribution of the partial correlation coefficient. Metron, 3, 329-332.

Galton, F. (1889). Natural Inheritance. London: Macmillan.

Heck, D. L. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root. Annals of Mathematical Statistics, 31, 625-642.

Hotelling, H. (1935) The most predictable criterion. Journal of Educational Psychology, 26, 139-142.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28, 321-377.

McNemar, Q. (1969). Psychological Statistics, Fourth Edition, New York: John Wiley.

Morrison, D. F. (1967). Multivariate Statistical Methods. New York: McGraw Hill.

Pearson, K. (1896). Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. Philosophical Transactions of the Royal Society of London, Series A, 187, 253-318.

Pearson, K. (1898). Mathematical contributions to the theory of evolution V. On the reconstruction of the stature of prehistoric races. Philosophical Transactions of the Royal Society of London Series A, 192, 169-244.

Rao, B. R. (1969). Partial Canonical Correlacions, Trabajos de Estadistica y de Investigacion operativa, XX, 211-219.

Rao, C. R. (1952). Advanced statistical methods in biometric research. New York: John Wiley.

Rao, C. R. (1973). Linear Statistical Inference and its applications, Second Edition. New York: John Wiley.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics, 24, 220-238.

Roy, S. N. (1957). Some Aspects of Multivariate Analysis. New York: John Wiley.

Stewart, D. K. and W. A. Love (1968). A general canonical correlation index. Psychological Bulletin, 70, 160-163.

Timm, N. H. (1975). Multivariate Analysis with applications in Education and Psychology. Belmont: Brooks - Cole.

Williams, E. J. (1967). The analysis of association among many variables. Journal of the Royal Statistical Society, Series B, 29, 199-242.

Yule, G. U. (1897). On the theory of correlation. Journal of the Royal Statistical Society, 60, 812-854.

Yule, G. U. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. Proceedings of the Royal Society of London, Series A, 79, 182-193.